Info 206: Computing

Lecture 6

Arrays and Lists

September 8, 2015

How are characters & strings sorted?

- Depends on representation
  - ASCII (important for us)
  - EBCDIC (rare)
  - Unicode (important for Web)

The Problem

- Representing text strings, such as `Hello, world`, in a computer

Codes and Characters

- Each character is coded as a byte
- Most common coding system is ASCII
- ASCII = American National Standard Code for Information Interchange
- Defined in ANSI document X3.4-1977

ASCII Features

- 7-bit code
- 8^{th} bit is unused
- $2^7 = 128$ codes
- Two general types of codes:
  - 95 are “Graphic” codes (displayable on a console)
  - 33 are “Control” codes (control features of the console or communications channel)

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	[illegible]	1	A	Q	a	q
0010	STX	DC2	[illegible]	2	B	R	b	r
0011	ETX	DC[illegible]	[illegible]	[illegible]	[illegible]	S	c	s
0100	EDT	DC[illegible]	[illegible]	[illegible]	[illegible]	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
011[illegible]	BEL	ETB	'	7	G	W	g	w
100[illegible]	BS	CAN	(	8	H	X	h	x
100[illegible]	HT	EM	)	9	I	Y	i	y
101[illegible]	LF	SUB	*	:	J	Z	j	z
[illegible]	[illegible]	[illegible]	+	;	K	[	k	{
[illegible]	[illegible]	[illegible]	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Most significant bit

Least significant bit

e.g., ‘a’ = 1100001

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

95 Graphic codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

33 Control codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Alphabetic codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Numeric codes

<table>
<tr><th></th><th>000</th><th>001</th><th>010</th><th>011</th><th>100</th><th>101</th><th>110</th><th>111</th></tr>
<tr><td>0000</td><td>NULL</td><td>DLE</td><td></td><td>0</td><td>@</td><td>P</td><td>`</td><td>p</td></tr>
<tr><td>0001</td><td>SOH</td><td>DC1</td><td>!</td><td>1</td><td>A</td><td>Q</td><td>a</td><td>q</td></tr>
<tr><td>0010</td><td>STX</td><td>DC2</td><td>"</td><td>2</td><td>B</td><td>R</td><td>b</td><td>r</td></tr>
<tr><td>0011</td><td>ETX</td><td>DC3</td><td>#</td><td>3</td><td>C</td><td>S</td><td>c</td><td>s</td></tr>
<tr><td>0100</td><td>EDT</td><td>DC4</td><td>$</td><td>4</td><td>D</td><td>T</td><td>d</td><td>t</td></tr>
<tr><td>0101</td><td>ENQ</td><td>NAK</td><td>%</td><td>5</td><td>E</td><td>U</td><td>e</td><td>u</td></tr>
<tr><td>0110</td><td>ACK</td><td>SYN</td><td>&</td><td>6</td><td>F</td><td>V</td><td>f</td><td>v</td></tr>
<tr><td>0111</td><td>BEL</td><td>ETB</td><td>'</td><td>7</td><td>G</td><td>W</td><td>g</td><td>w</td></tr>
<tr><td>1000</td><td>BS</td><td>CAN</td><td>(</td><td>8</td><td>H</td><td>X</td><td>h</td><td>x</td></tr>
<tr><td>1001</td><td>HT</td><td>EM</td><td>)</td><td>9</td><td>I</td><td>Y</td><td>i</td><td>y</td></tr>
<tr><td>1010</td><td>LF</td><td>SUB</td><td>*</td><td>:</td><td>J</td><td>Z</td><td>j</td><td>z</td></tr>
<tr><td>1011</td><td>VT</td><td>ESC</td><td>+</td><td>;</td><td>K</td><td>[</td><td>k</td><td>{</td></tr>
<tr><td>1100</td><td>FF</td><td>FS</td><td>,</td><td><</td><td>L</td><td>\</td><td>l</td><td>|</td></tr>
<tr><td>1101</td><td>CR</td><td>GS</td><td>-</td><td>=</td><td>M</td><td>]</td><td>m</td><td>}</td></tr>
<tr><td>1110</td><td>SO</td><td>RS</td><td>.</td><td>></td><td>N</td><td>^</td><td>n</td><td>~</td></tr>
<tr><td>1111</td><td>SI</td><td>US</td><td>/</td><td>?</td><td>O</td><td>_</td><td>o</td><td>DEL</td></tr>
</table>

Punctuation, etc.

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

“Hello, world” Example

		Binary		Hexadecimal		Decimal
H	=	01001000	=	48	=	72
e	=	01100101	=	65	=	101
l	=	01101100	=	6C	=	108
l	=	01101100	=	6C	=	108
o	=	01101111	=	6F	=	111
,	=	00101100	=	2C	=	44
	=	00100000	=	20	=	32
w	=	01110111	=	77	=	119
o	=	01100111	=	67	=	103
r	=	01110010	=	72	=	114
l	=	01101100	=	6C	=	108
d	=	01100100	=	64	=	100

Common Control Codes

• CR	0D	carriage return
• LF	0A	line feed
• HT	09	horizontal tab
• DEL	7F	delete
• NULL	00	null

Hexadecimal code

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Terminology

- Learn the names of the special symbols
  - [ ] brackets
  - { } braces
  - ( ) parentheses
  - @ ‘at’ sign
  - & ampersand
  - ~ tilde

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

EBCDIC

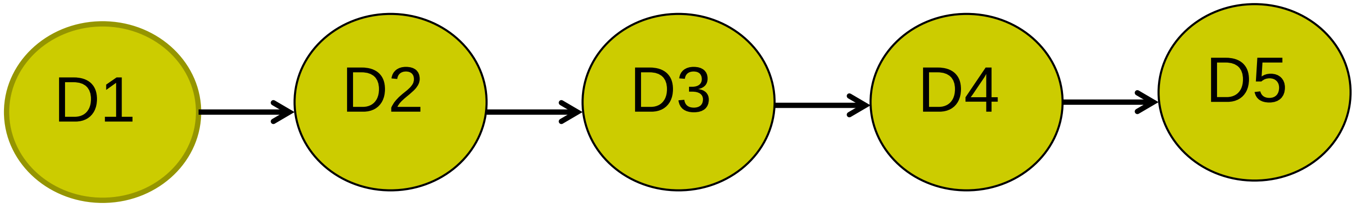
- Extended BCD Interchange Code
- 8-bit code
- Developed by IBM
- Rarely used today
- IBM mainframes only

Unicode

- Has over 110,000 different characters
- Multi-byte representation
- Developed by a consortium
- www.unicode.org

Linear Collections

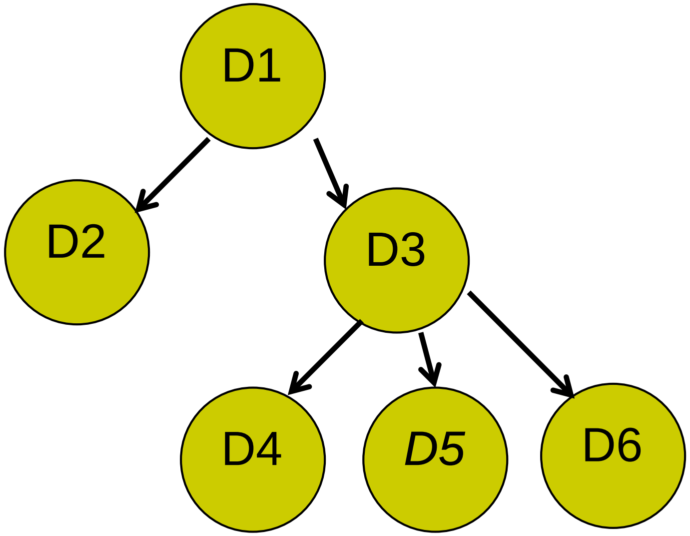
- Ordered by position



- Everyday examples:
  - Grocery lists
  - Stacks of dinner plates
  - A line of customers waiting at a bank

Hierarchical Collections

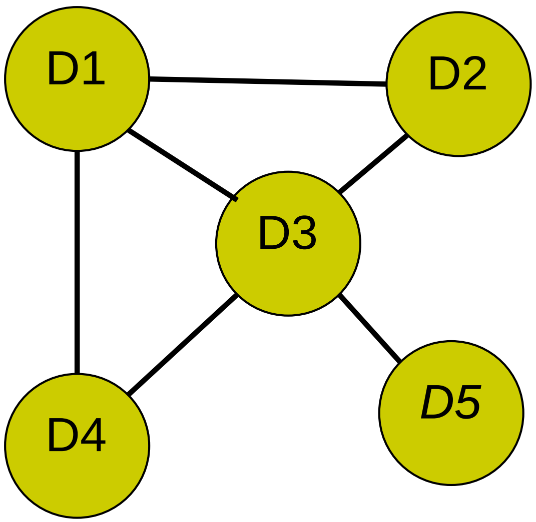
- Structure reminiscent of an upside-down tree



  - D3's **parent** is D1; its **children** are D4, D5, and D6
- Examples: a file directory system, a company's organizational tree, a book's table of contents

Graph Collections

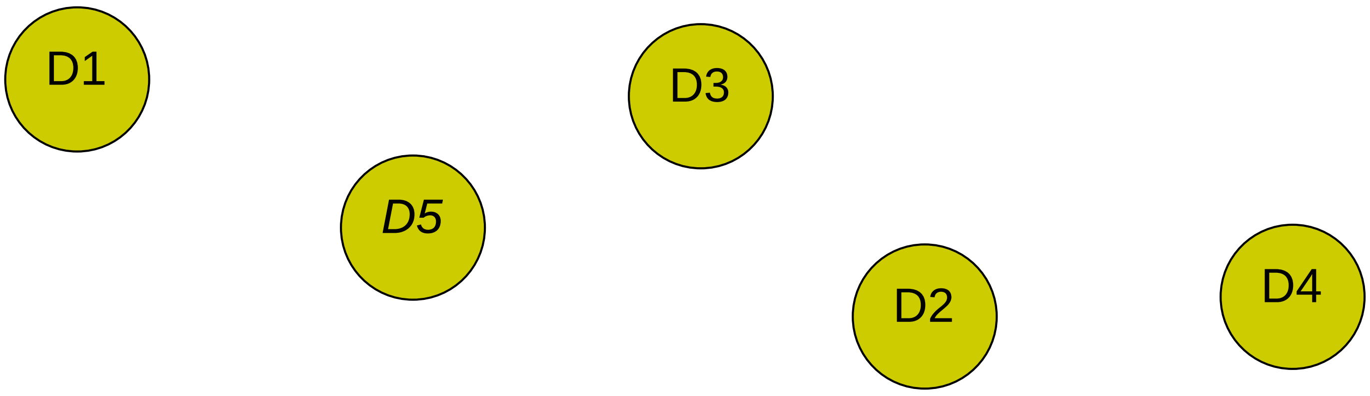
- **Graph**: Collection in which each data item can have many predecessors and many successors



  - D3's **neighbors** are its predecessors and successors
- Examples: Maps of airline routes between cities; electrical wiring diagrams for buildings

Unordered Collections

- Items are not in any particular order
    - One cannot meaningfully speak of an item’s predecessor or successor



- Example: Bag of marbles

Operations on Collections

- Search and retrieval
- Removal
- Insertion
- Replacement (removal/insertion)
- Traversal
    - If we can traverse with Python `for` loop, then *iterable*

Operations on Collections

- Tests for equality
  - On elements in a collection
  - On the collection as a total
- Determine size
  - Some collections may have maximum capacity
- Cloning
  - Sometimes clone collections contain the same items
  - Deep copy: clone both collection and items

Abstraction and Abstract Data Types

- To a user, a collection is an abstraction
- In CS, collections are **abstract data types (ADTs)**
  - ADT users are concerned with learning its interface
  - Developers are concerned with implementing their behavior in the most efficient manner possible
- In Python, methods are the smallest unit of abstraction, classes are the next in size, and modules are the largest
- We will implement ADTs as classes or sets of related classes in modules

Data Structures for Implementing Collections: Arrays

- “Data structure” and “**concrete data type**” refer to the internal representation of an ADT’s data
- The two data structures most often used to implement collections in most programming languages are **arrays** and **linked structures**
  - Different approaches to storing and accessing data in the computer’s memory
  - Different space/time trade-offs in the algorithms that manipulate the collections

Python hides data organization

- Lists (arrays – but they can grow)
- Tuples (array on heap)
- Dictionaries (hash tables)

- But it is there “under the covers”
- Sometimes we need to more explicitly address data organization

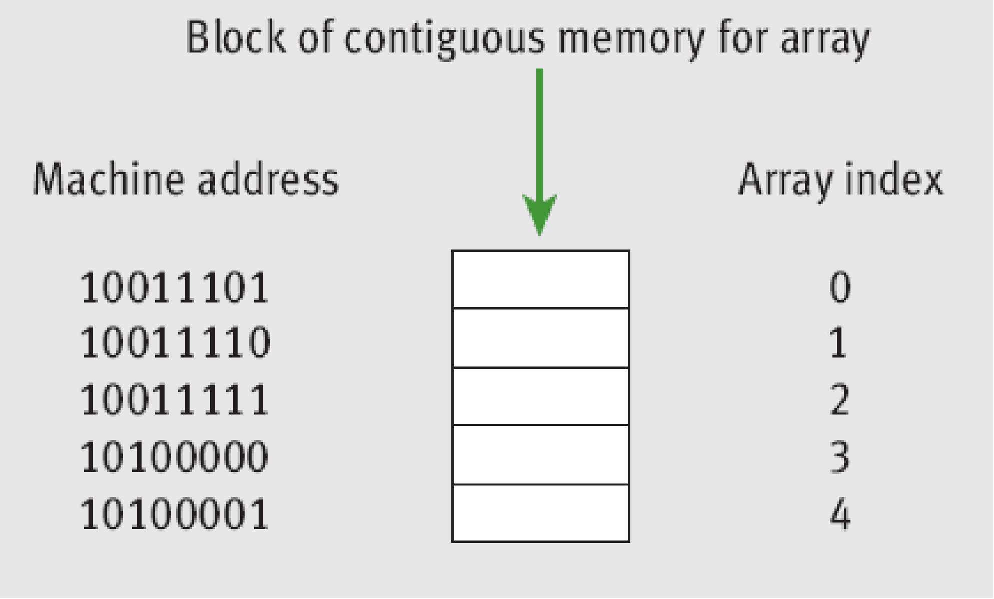
The Array Data Structure

- Array: Underlying data structure of a Python list
  - More restrictive than Python lists
- We'll define an **`Array`** class

User's Array Operation	Method in the Array Class
`a = Array(10)`	`__init__(capacity, fillValue=None)`
`len(a)`	`__len__()`
`str(a)`	`__str__()`
`for item in a:`	`__iter__()`
`a[index]`	`__getitem__(index)`
`a[index] = newitem`	`__setitem__(index, newItem)`

Random Access and Contiguous Memory

- Array indexing is a **random access** operation



- Address of an item: **base address** + **offset**
    - Index operation has two steps:
        - Fetch the base address of the array's memory block
        - Return the result of adding the index*k to this address

Static and Dynamic Arrays

- Arrays in older languages were static
- Modern languages support **dynamic arrays**
- To readjust length of an array at run time:
  - Create an array with a reasonable default size at start-up
  - When it cannot hold more data, create a new, larger array and transfer the data items from the old array
  - When the array seems to be wasting memory, decrease its length in a similar manner
- These adjustments are automatic with Python lists